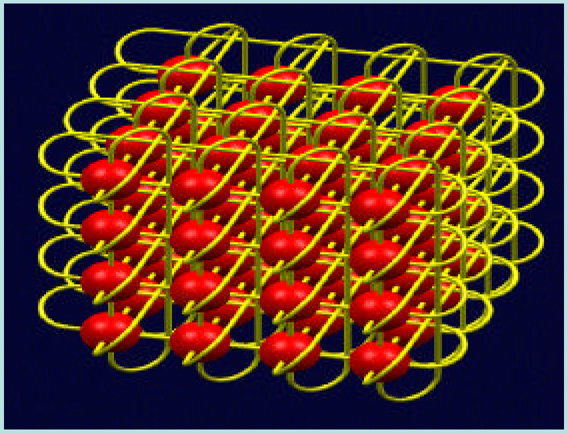


BlueGene/L System Package

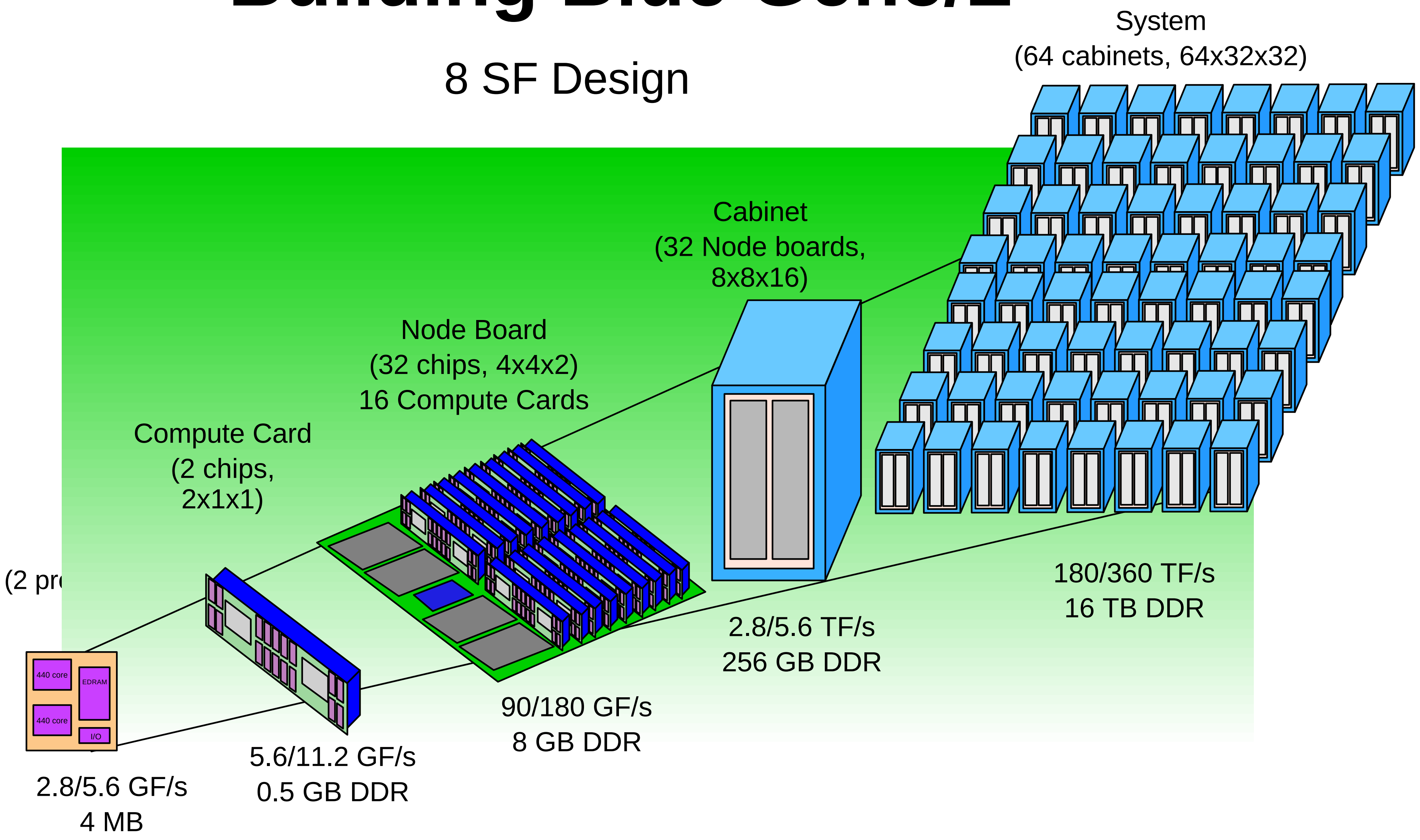
Update for BG/L Tahoe Conference

8/13/2002

Paul Coteus IBM

Building Blue Gene/L

8 SF Design



(2 pr

2.8/5.6 GF/s
4 MB

Compute Card
(2 chips,
2x1x1)

5.6/11.2 GF/s
0.5 GB DDR

Node Board
(32 chips, 4x4x2)
16 Compute Cards

90/180 GF/s
8 GB DDR

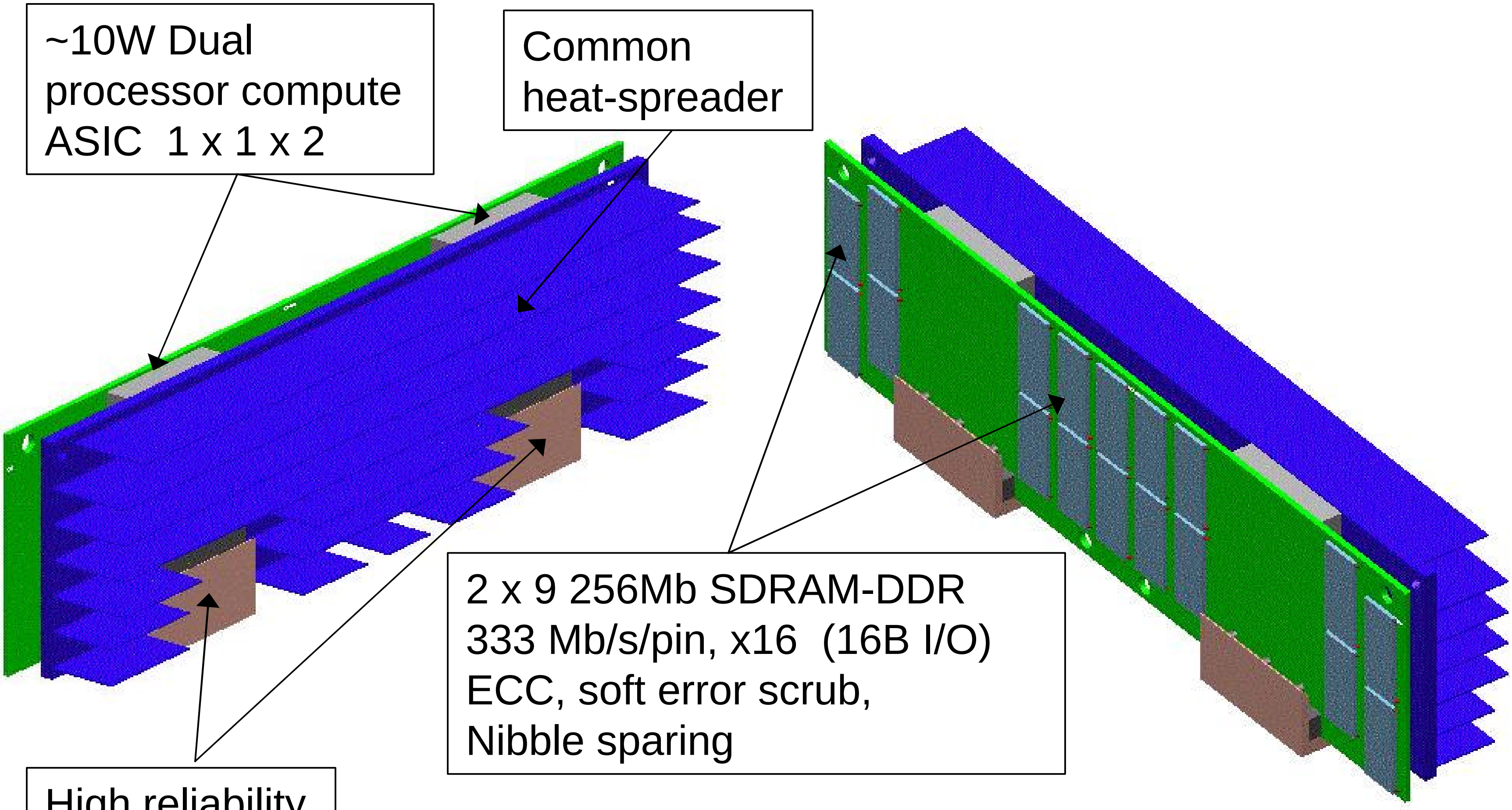
Cabinet
(32 Node boards,
8x8x16)

2.8/5.6 TF/s
256 GB DDR

System
(64 cabinets, 64x32x32)

180/360 TF/s
16 TB DDR

BG/L Dual Compute Card



~10W Dual processor compute ASIC 1 x 1 x 2

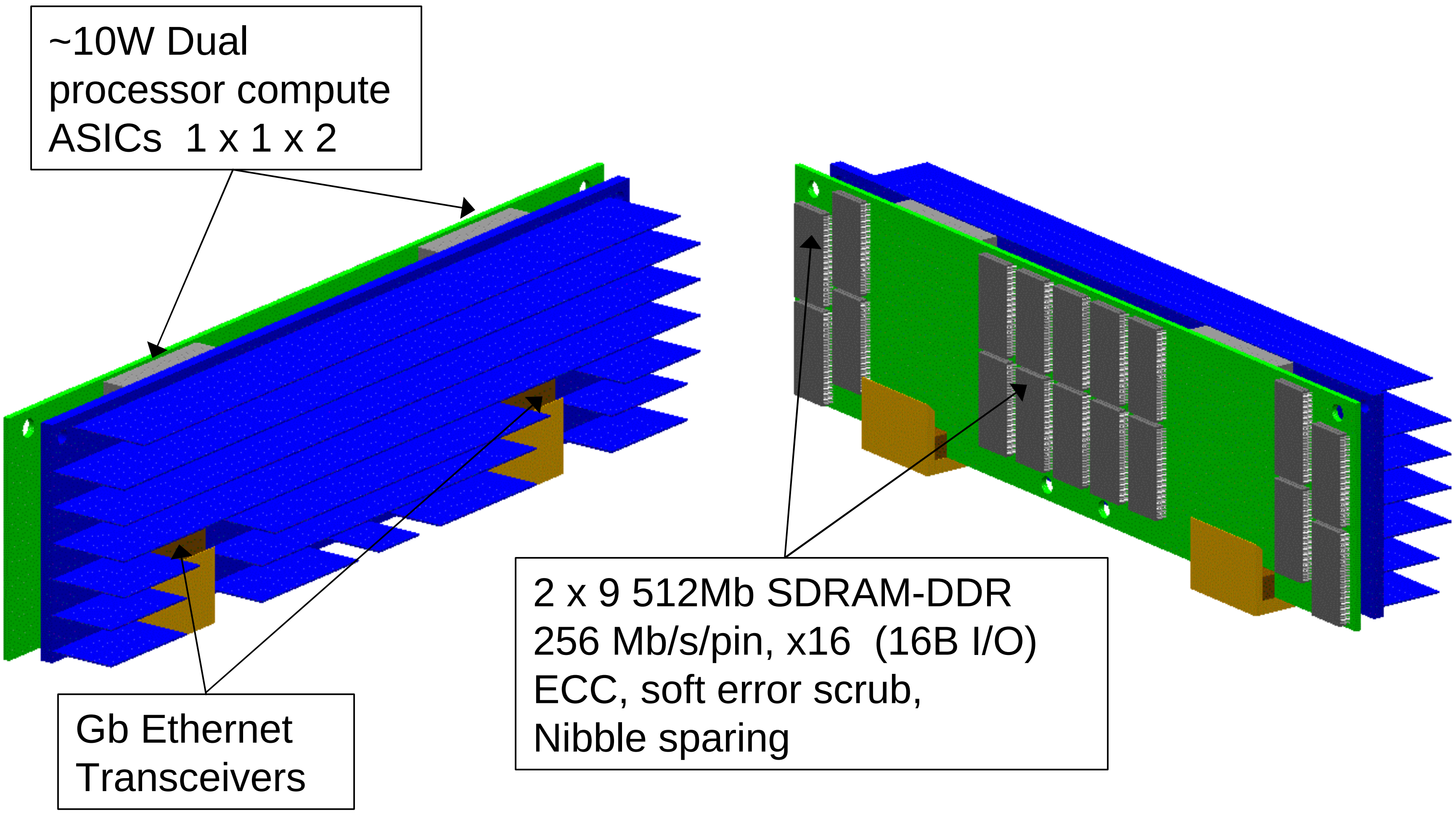
Common heat-spreader

High reliability connectors

2 x 9 256Mb SDRAM-DDR
333 Mb/s/pin, x16 (16B I/O)
ECC, soft error scrub,
Nibble sparing

~200mm x 50mm x 30mm pitch
~30W max

BG/L Dual IO Card



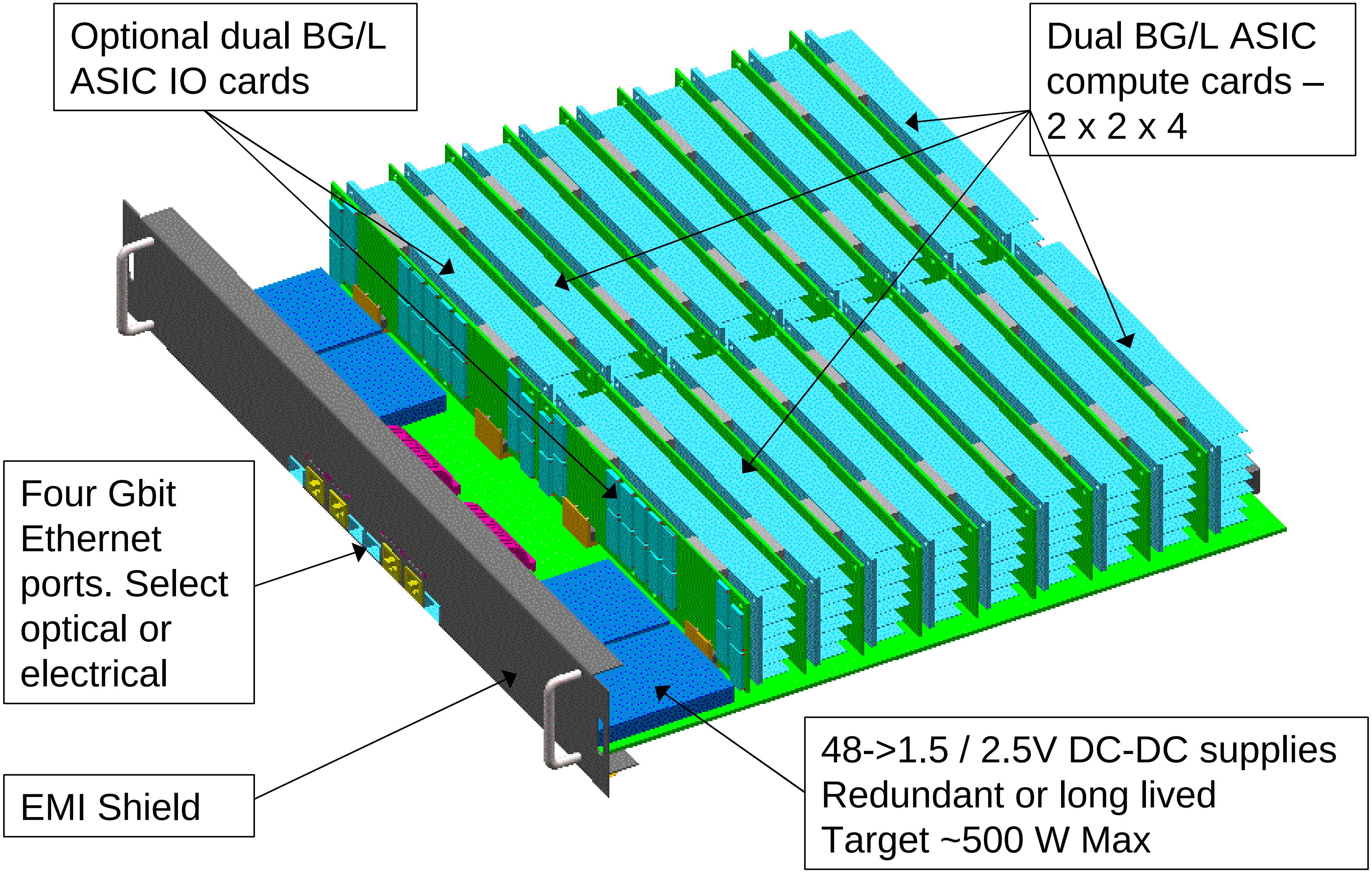
~10W Dual processor compute ASICs 1 x 1 x 2

Gb Ethernet Transceivers

2 x 9 512Mb SDRAM-DDR
256 Mb/s/pin, x16 (16B I/O)
ECC, soft error scrub,
Nibble sparing

~200mm x 50mm x 30mm pitch
~30W max

BG/L 32-way node card



Optional dual BG/L ASIC IO cards

Dual BG/L ASIC compute cards – 2 x 2 x 4

Four Gbit Ethernet ports. Select optical or electrical

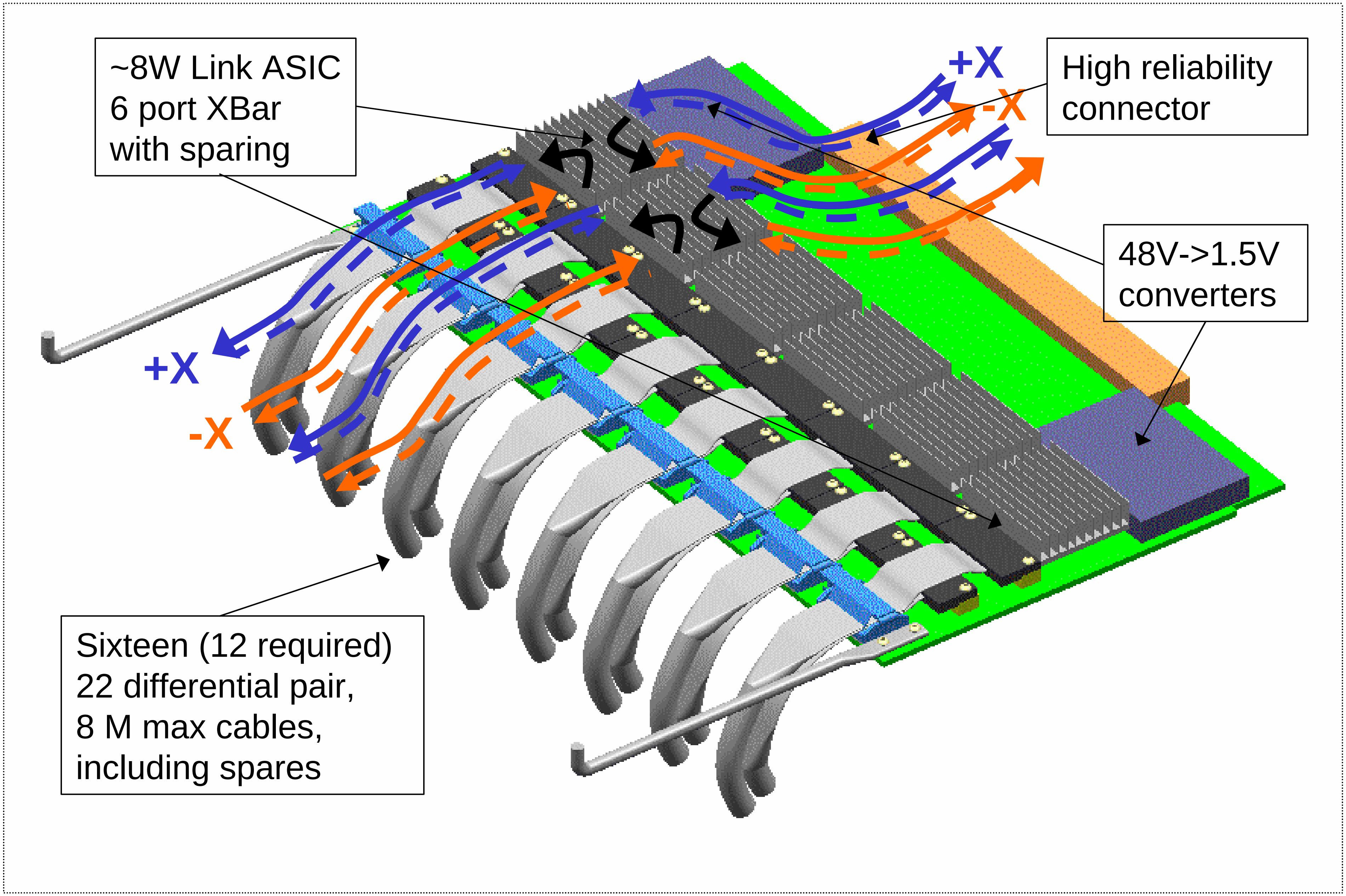
EMI Shield

48->1.5 / 2.5V DC-DC supplies
Redundant or long lived
Target ~500 W Max

BG/L Link Card

- Redrives 1.4Gb/s serial torus & tree links between midplanes
- Redirects links to skip over failed midplanes
- Creates electrically isolated independent machine sub-partitions

BG/L Link Card



~8W Link ASIC
6 port XBar
with sparing

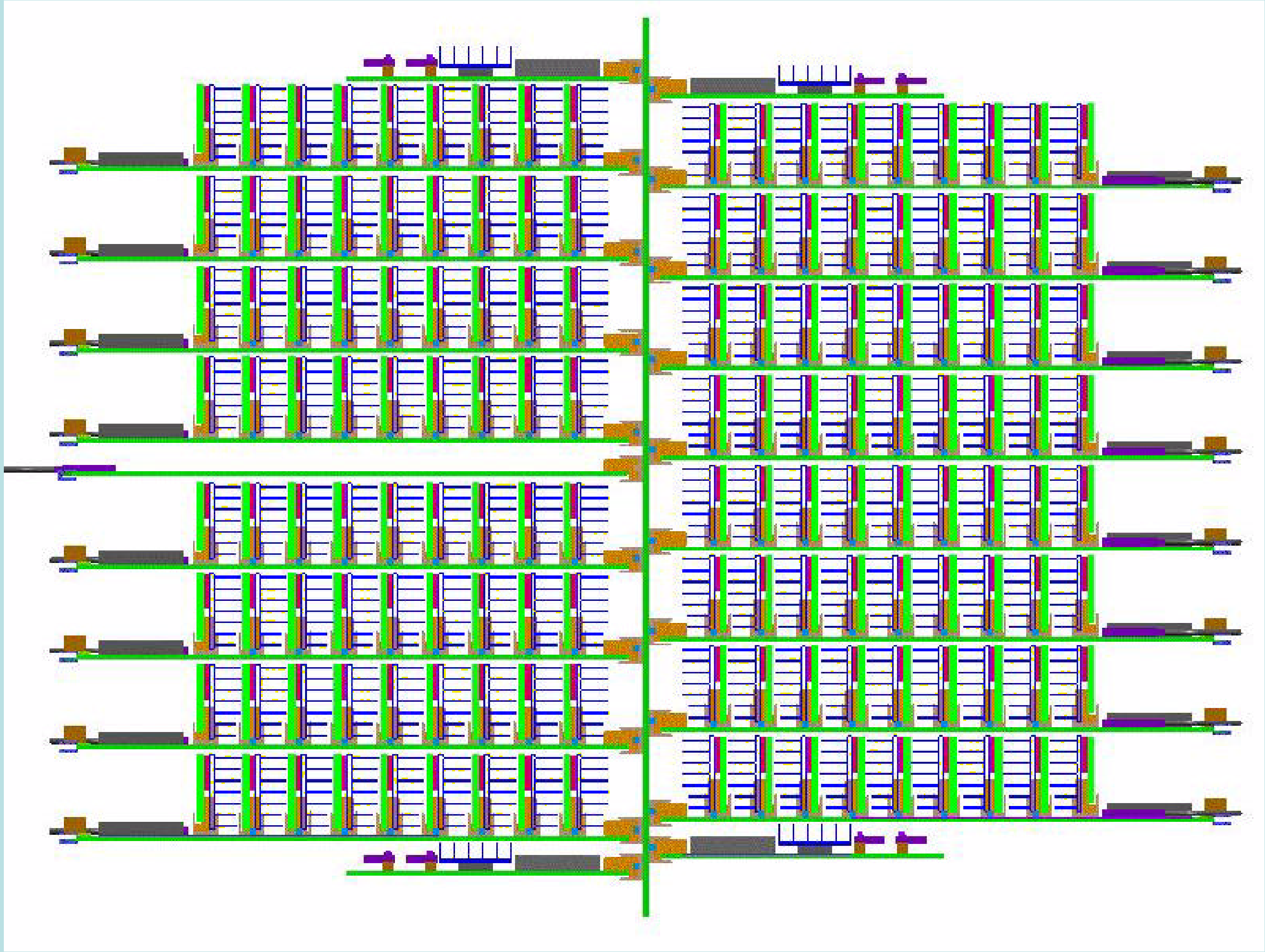
High reliability
connector

48V->1.5V
converters

Sixteen (12 required)
22 differential pair,
8 M max cables,
including spares

BG/L Midplane

- 8 x 8 x 8 torus / tree subpartiton
- ~9 KW max inc fans
- 4 dual Gigabit Enet I/O cards
- 16 compute cards
- 4 link cards
- Service card (clock, control, persistent power)
- Distributes 48V DC, fan power, service port



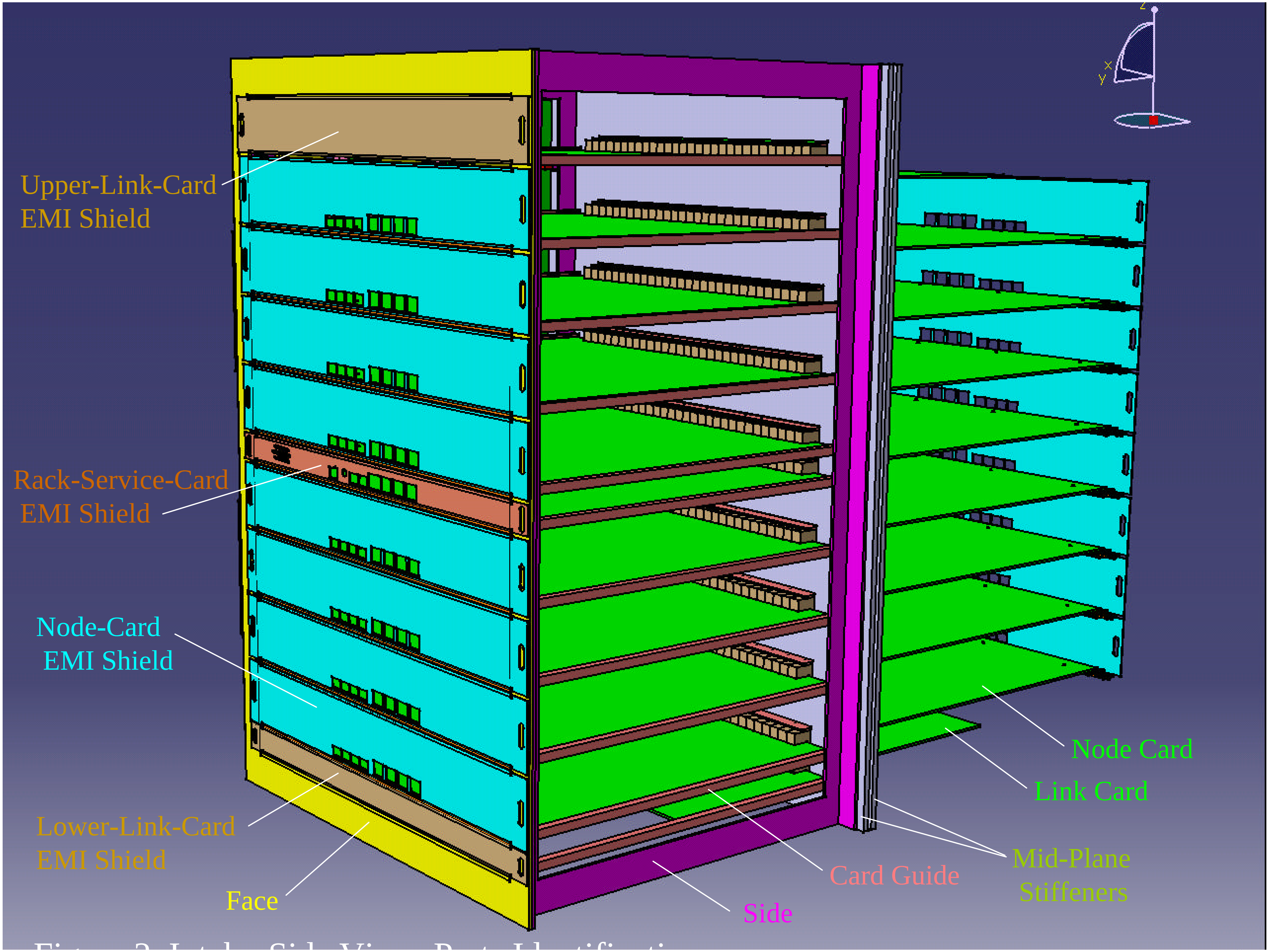


Figure 3: Latch Side View Parts Identification

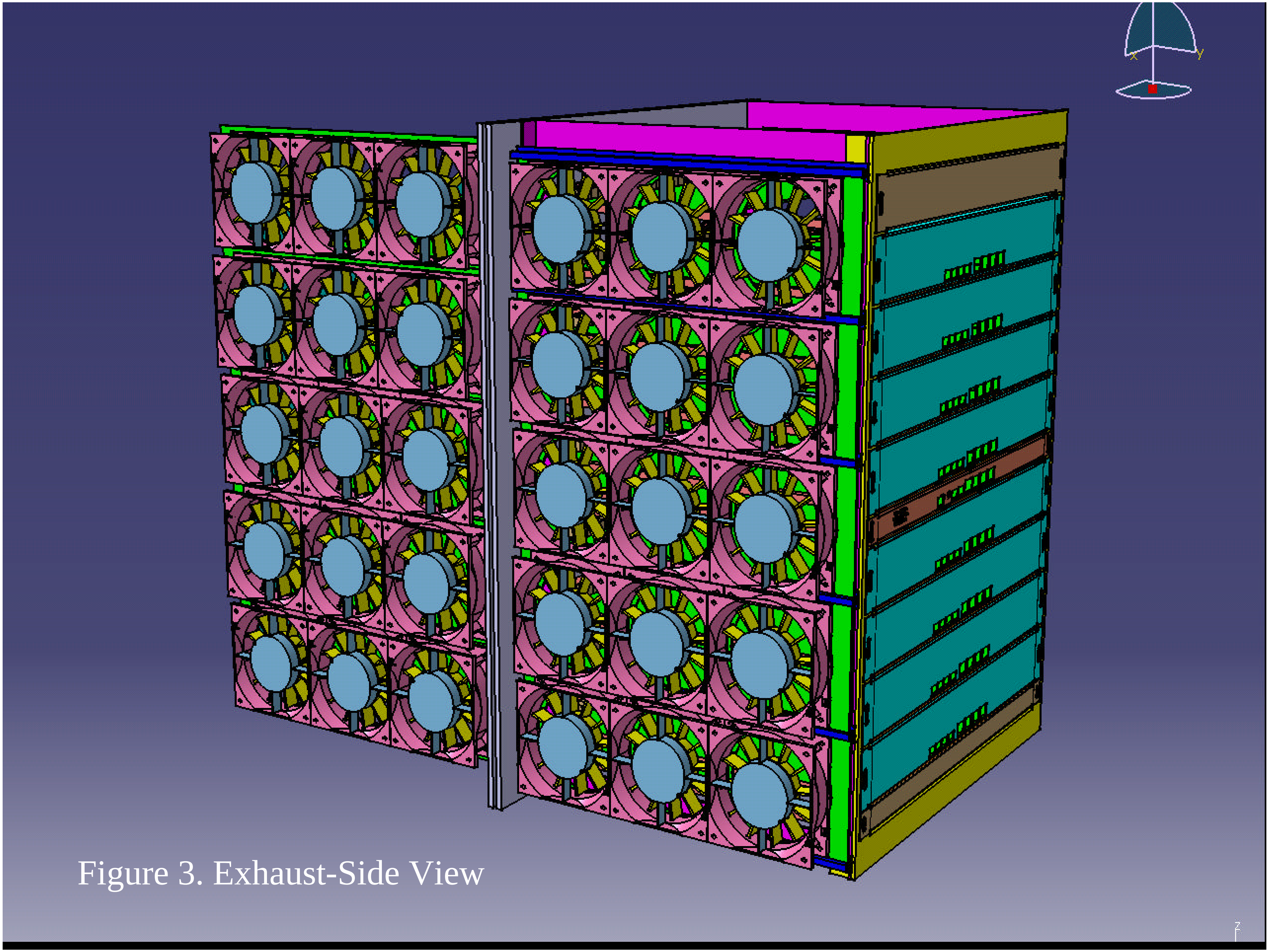


Figure 3. Exhaust-Side View

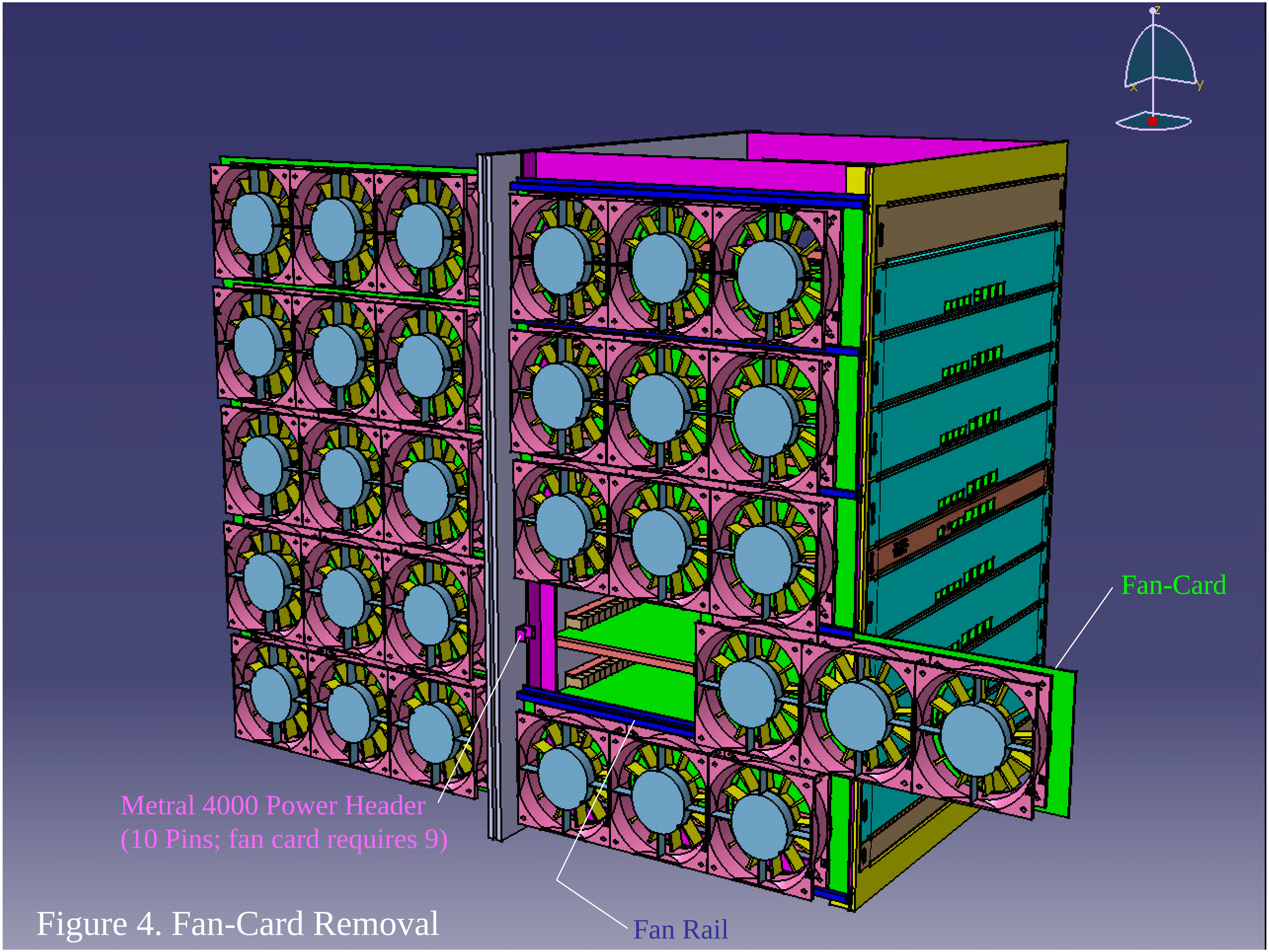
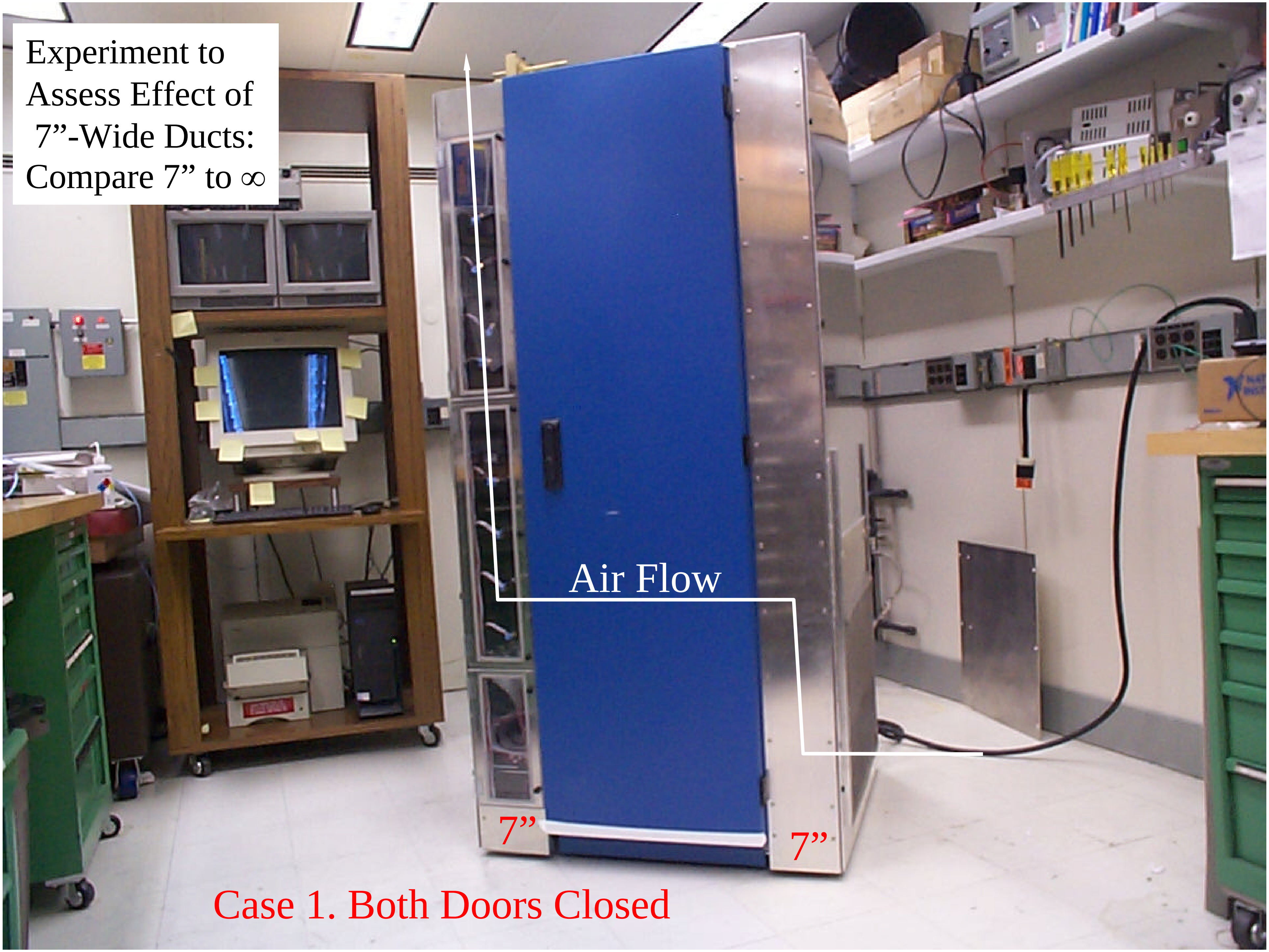


Figure 4. Fan-Card Removal

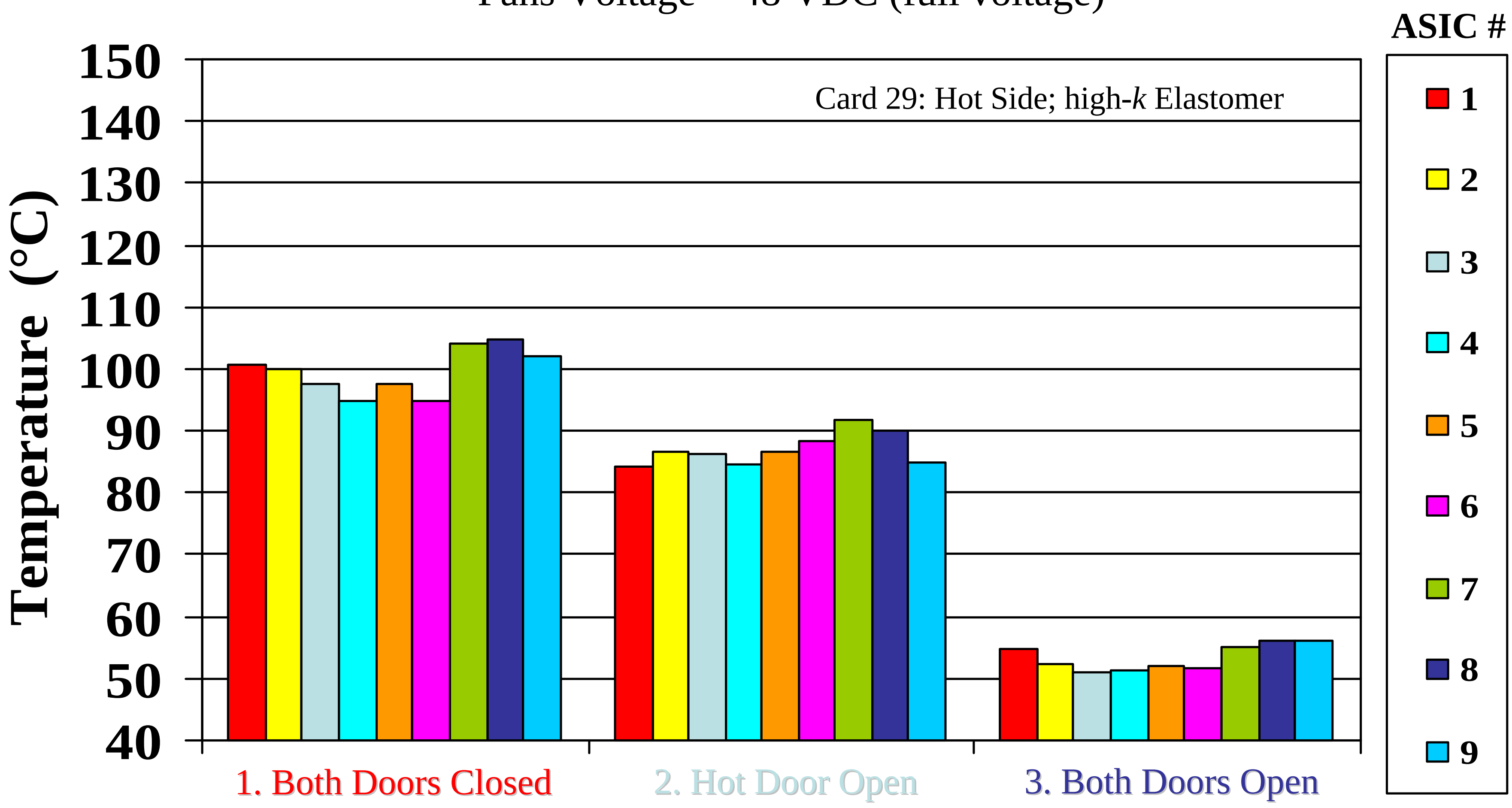
Experiment to Assess Effect of 7"-Wide Ducts: Compare 7" to ∞



Case 1. Both Doors Closed

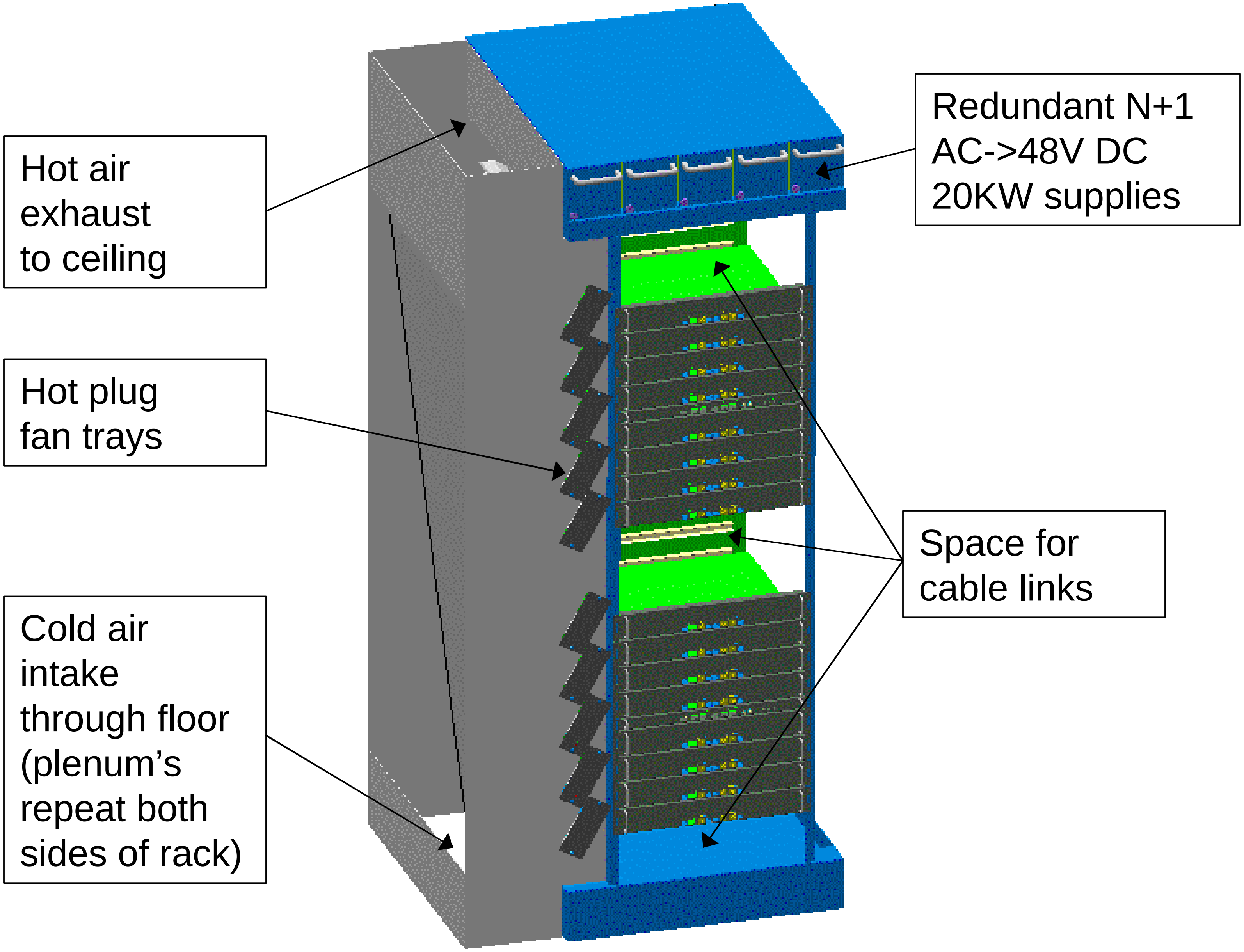
Equilibrium Temperatures on Card 29 vs. Door Configuration

Air-Inlet Temperature = 27 °C (7 °C hotter than in real machine)
Fans Voltage = 48 VDC (full voltage)



Shawn Hall 4-01-02
Run: 02-04-01 07.48 Equilibrium T vs. Door Configuration, V=48

BG/L Rack – Covers Off - Concept



Hot air exhaust to ceiling

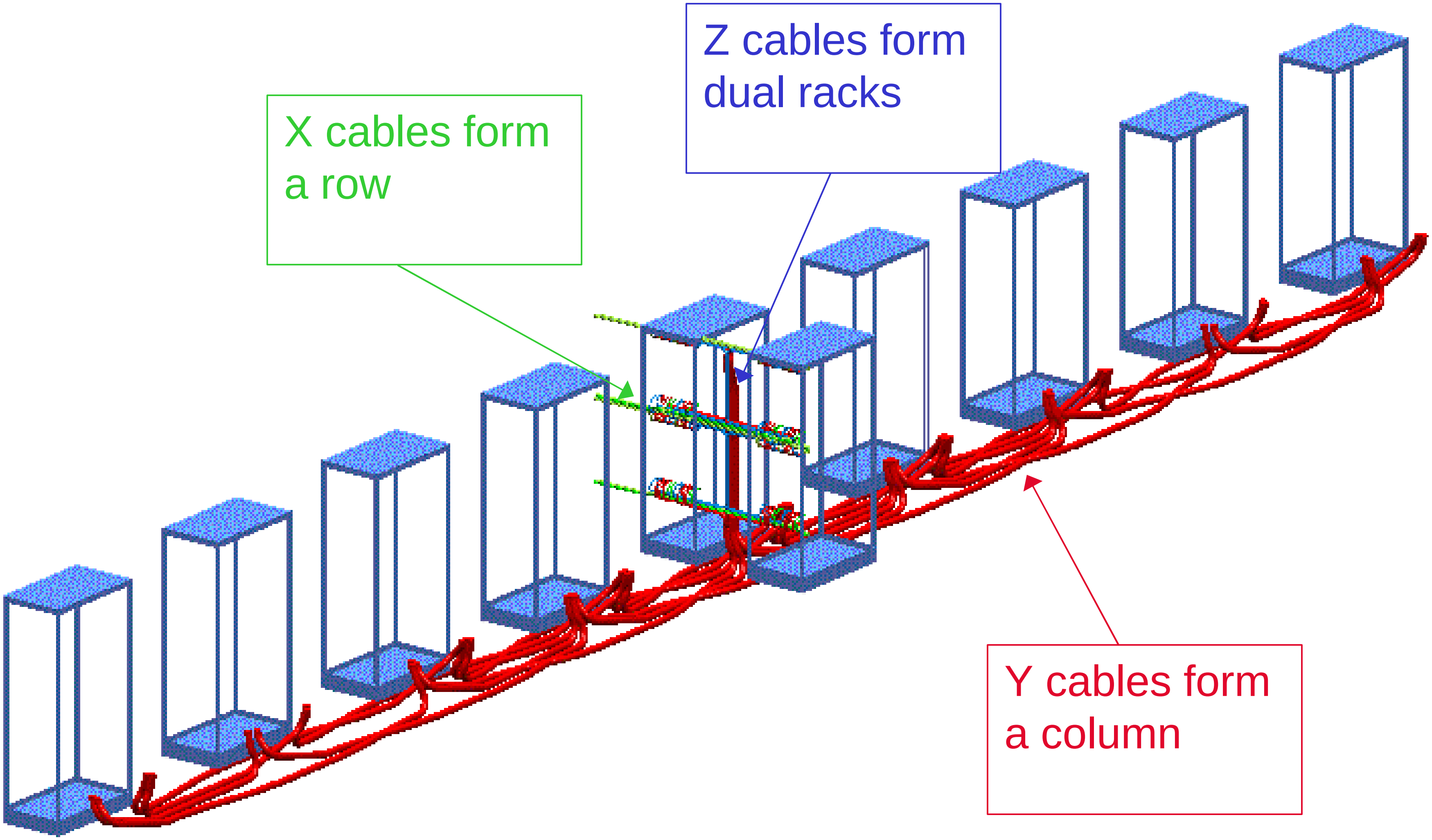
Hot plug fan trays

Cold air intake through floor (plenum’s repeat both sides of rack)

Redundant N+1 AC->48V DC 20KW supplies

Space for cable links

BG/L Racks with X-Y-Z Torus Cables

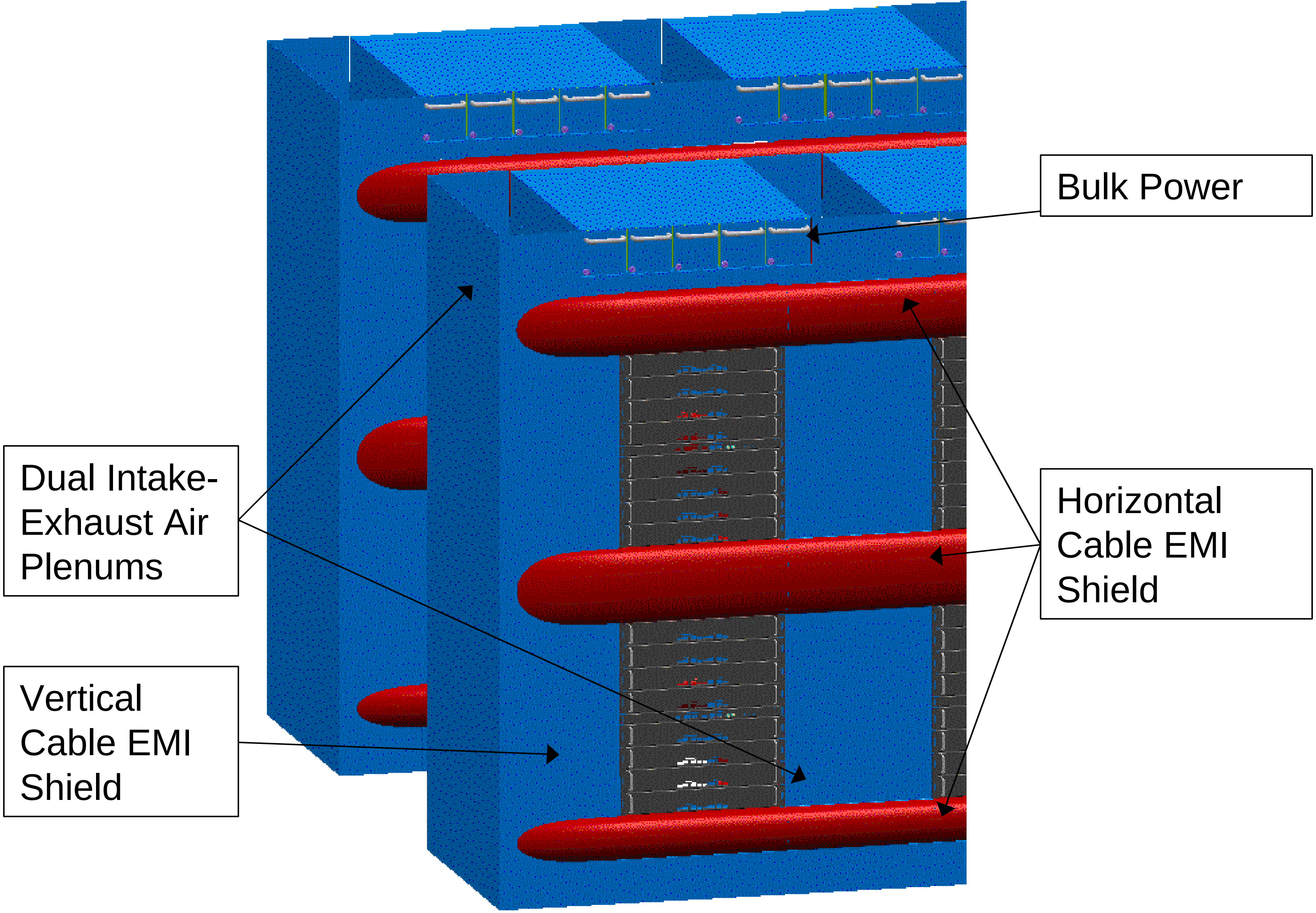


X cables form a row

Z cables form dual racks

Y cables form a column

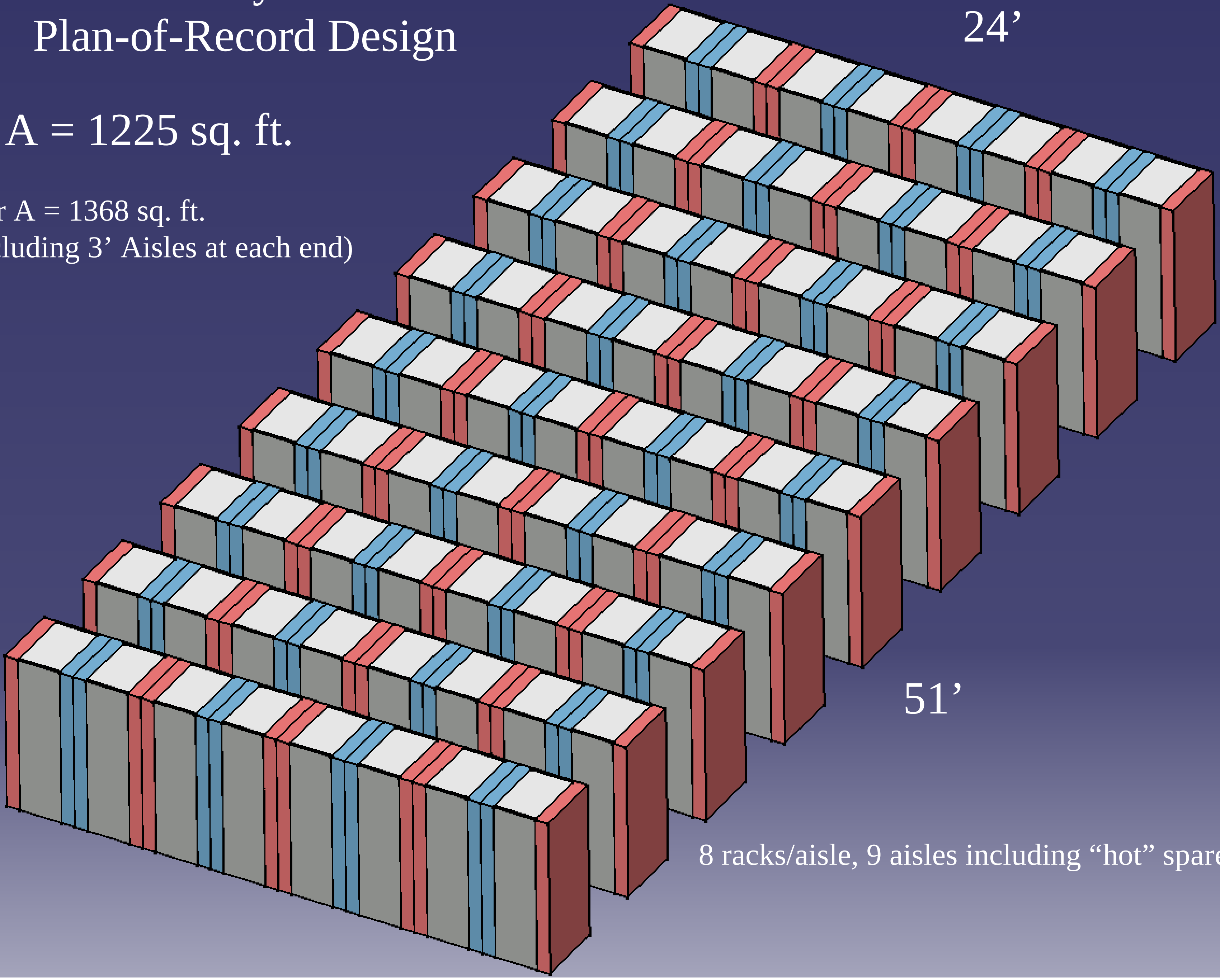
Racks with Covers - Concept


Bulk Power
Dual Intake-Exhaust Air Plenums
Horizontal Cable EMI Shield
Vertical Cable EMI Shield

Livermore System:
Plan-of-Record Design

A = 1225 sq. ft.

(or A = 1368 sq. ft.
Including 3’ Aisles at each end)



8 racks/aisle, 9 aisles including “hot” spare

BG/L Reliability & Serviceability

- Individually & globally addressable JTAG network to all ASICs through chainable Gb Ethernet service port.

- ECC or parity / retry with sparing on most busses.

- Uncorrectable errors cause restart from checkpoint.

- Most hardware fails not covered by redundancy or sparing are repaired by switching to “hot” spare racks.

- Only fails early in global clock tree, or certain failures of link cards, require immediate service.

BG/L Reliability Estimates

Component	FIT per component*	Components per 64k partition	FITs per system	Failure rate per week
ETH complex	160	2704	433k	
DRAM	5	599,040	2,995k	
Compute + I/O ASIC	20	66,560	1,331k	
Link ASIC	10	3072	60k	
Clock chip	6.5	~1200	8k	
Non-redundant power supply	500	384	384k	
Total (65,536 compute nodes)			5211k	0.87

•After burn-in and applied redundancy.
T=60C, V=Nom, 40K POH.
FIT = Failures in parts per million per thousand power-on hours.
1 FIT = 0.168*10^{-6} fails/week if the machine runs 24 hrs/day.